基于最优运输理论的物联网边缘计算资源优化机制

张琪[1]，蒋宇娜[1]，葛晓虎[1]，李永会[2]
（1. 华中科技大学，湖北 武汉 430074；2. 澳大利亚悉尼大学，澳大利亚 悉尼 NSW）

摘　要：随着物联网和边缘计算的发展，物联网设备可以将计算密集型任务卸载到边缘计算服务器上进行处理。由于物联网设备分布以及计算需求的变化，需要对边缘计算资源进行动态管理。利用最优运输理论对物联网中计算资源分配进行优化，提出一种基于物联网设备分布和边缘计算服务器位置的区域优化划分机制，在边缘计算服务器计算能力的约束下对物联网设备的能耗以及时延性能进行优化。仿真结果表明，与传统泰森多边形划分机制相比，该优化机制有更好的均衡性，并且物联网设备的平均能耗最多降低 21%，平均时延最多降低 45%。

关键词：物联网；边缘计算；资源分配；最优运输理论；能耗；时延
中图分类号：TN92
文献标识码：A
doi: 10.11959/j.issn.2096−3750.2021.00225

Resource allocation based on optimal transport theory in IoT edge computing

ZHANG Qi[1], JIANG Yuna[1], GE Xiaohu[1], LI Yonghui[2]
1. Huazhong University of Science and Technology, Wuhan 430074, China
2. The University of Sydney, Sydney NSW, Australia

Abstract: With the development of the Internet of things (IoT) and edge computing, the computation-intensive tasks of IoT devices can be offloaded to edge devices and processed at the edge of networks. Due to the variation of the distribution and computation requirements of IoT devices, the computation resources of edge networks need to be managed dynamically. The optimal transport theory was adopted to optimize the computation resources allocation in IoT networks. An optimized regional partition mechanism was proposed based on the distribution of IoT devices and locations of edge computing devices. Under constraints on the computing capabilities of edge computing devices, the energy consumption and delay of IoT devices were optimized. The simulation results show that, compared with the traditional Voronoi partition scheme, the proposed optimization mechanism shows better balance. The average transmitting power can be reduced by 21% and the average delay can be reduced by 45%.

Key words: Internet of things, edge computing, resource allocation, optimal transport theory, energy consumption, delay

1　引言

近年来，物联网设备被用于医疗、工业、交通等领域[1-2]。物联网设备的日益普及正在推动计算密集型移动应用程序的发展，然而物联网设备能力有限，很难在本地处理复杂的计算任务。云计算将计算任务从设备端转移到公共云中，从而可以缓解计算密集型应用程序和资源有限的物联网设备之间的矛盾关系。然而传统的云计算系统依赖于远程的公共云，需要远程的数据交换。物联网设备产生的大量数据可能会阻塞移动通信网络，而且在远程物联网云中处理传感数据可能无法确保应用的时延

收稿日期：2021−01−04；**修回日期：**2021−02−03
通信作者：葛晓虎，xhge@mail.hust.edu.cn
基金项目：国家自然科学基金资助项目（No.U2001210）
Foundation Item: The National Natural Science Foundation of China (No.U2001210)

需求。与云计算不同的是，边缘计算在无线接入网络中提供计算能力，使得多种应用程序和服务能够在网络边缘运行[3]。边缘计算通过将物联网设备上的计算任务卸载到邻近的边缘计算服务器上，将物联网设备从计算密集型的工作负载中解放出来。由于物联网设备分布以及计算需求的变化，需要对计算资源进行动态管理，以适应时变计算需求。针对时变的物联网设备空间分布以及计算需求，如何对计算资源进行合理分配以达到最优的性能仍是亟待解决的问题。

为了优化能耗[4-8]和时延[9-10]性能，并且对能耗和时延做出权衡[10-13]，大量文献对边缘计算卸载过程、计算过程进行研究。文献[4]专注于平衡本地计算能耗与卸载能耗，提出了最优卸载决策策略。文献[5]研究了基于时分多址和正交频分多址的多用户边缘计算系统的资源分配问题，并且针对容量有限的云，提出了一种次优资源分配算法。文献[6]使用博弈论设计了单云多用户的分布式计算卸载，以实现能量和时延的最小化。在存在时延约束的条件下，文献[7]综合考虑了无线资源和计算资源的联合分配，以达到能耗最小化。文献[8]研究了在中心云与边缘计算并存的情况下，用户选择不同云的调度问题。文献[14]通过计算卸载决策和计算资源配置的联合优化，提出了云计算与边缘计算协同情况下的计算卸载以及计算资源分配方案。文献[9]采用马尔可夫决策过程方法处理双时间尺度随机优化问题，根据任务缓冲区的排队状态、本地处理单元的执行状态以及传输单元的状态来调度计算任务。文献[10]提出了一种在边缘计算平台上分配计算资源的新模型，该模型允许服务提供商与边缘基础设施提供商预先建立资源共享契约。基于已建立的契约，服务提供商采用一种感知时延的调度和资源供应算法，使任务能够完成并满足它们的时延要求。在文献[11-12]中，利用 Lyapunov 优化技术分别研究了异构计算类型和多核移动设备云计算系统中能量和时延的权衡问题。类似的，在文献[13]中，作者提出了一个利用能耗和时延之间的权衡来联合优化无线电和计算资源使用的框架。文献[15]基于边缘计算过程中边缘和核心之间的相互作用和合作，研究了边缘计算系统中能耗和传输时延之间的权衡问题。深度强化学习在边缘计算资源分配方面得到了广泛应用。文献[16]将 Q 函数分解技术与双深 Q 网络相结合，提出了一种求解随机计算卸载问题的学习算法。文献[17]提出了一种空间−空中−地面综合网络边缘计算架构，针对此架构提出了一种基于深度强化学习的计算卸载方法。为自适应地分配计算和网络资源，减少平均服务时间，文献[18]提出了一种基于智能深度强化学习的资源分配方案。但是，采用深度强化学习的方法进行资源分配需要大量的数据进行训练，并且计算复杂度较高。目前，在对边缘计算资源分配机制的研究中，一般假设物联网设备随机分布在一定区域内，没有讨论在物联网设备分布不均匀的情况下边缘计算的资源分配机制。在物联网设备分布不均匀的情况下，物联网设备在一定区域的密集分布将造成部分边缘计算服务器任务过重或计算资源浪费。

最优运输理论是由 Monge 问题发展出来的理论体系[19]。最优运输理论在经济学、自动化控制、无线网络和人工智能等众多领域中都有运用。在无线网络方面，最优运输理论在资源匹配中得到了广泛的应用，通过资源与目标的匹配提高资源使用效率[20-22]。文献[23]研究了异构网络下的移动用户关联问题，它应用最优运输理论确定每个基站对应的小区，使正交频分多址蜂窝网络的总传输功率消耗最小化。在文献[24]中，作者研究了一个包含无人机基站和无人机用户的三维蜂窝网络，并且利用最优运输理论，将单元划分问题建模为易于求解的半离散最优运输问题，提出了一种基于公平资源分配方案的最优三维单元划分方法，使数据服务总量达到最大。文献[25]针对多架无人机作为飞行基站的最优部署问题进行了更深入的研究，利用最优运输理论确定了无人机的最佳位置及其覆盖区域的单元边界。然而，最优运输理论在基于边缘计算的物联网资源分配中仍没有得到充分的研究。

为提高物联网边缘计算场景中计算资源的使用效率、优化能耗以及时延性能，本文基于最优运输理论对物联网计算资源进行优化，主要贡献如下。

1) 基于 Kantorovich 对偶定理将物联网边缘计算系统中能耗和时延优化问题转化为可解的 Monge-Kantorovich 问题，进而利用梯度下降法给出了最优解。

2) 考虑物联网设备分布不均匀的情景，本文提出了一种基于最优运输理论的计算资源分配机制，解决了在边缘计算服务器计算能力有限且相同的

约束下，计算资源的分配问题，实现能耗和时延的优化。

3) 仿真分析了基于最优运输理论计算资源分配机制的性能。与泰森多边形划分机制相比，本文提出的新机制可使物联网设备的平均能耗最多降低 21%，平均时延最多降低 45%；与随机分配机制相比，本文提出的新机制可使物联网设备的平均能耗最多降低 24%，平均时延最多降低 51%。

2 系统建模

2.1 系统模型

本文考虑物联网中的低能耗设备，如传感器等不具备处理复杂任务能力的设备，它们需要将计算密集型任务卸载到边缘计算服务器进行处理，系统模型如图 1 所示。物联网设备与边缘计算服务器分布在区域 $\mathcal{Z}$ 中，其中 $\mathcal{Z}\subset\mathbb{R}^2$，$\mathbb{R}^2$ 代表二维空间。区域 $\mathcal{Z}$ 被分割为 M 个部分，表示为 $\mathcal{V}_1\cup\mathcal{V}_2\cup\cdots\cup\mathcal{V}_M=\mathcal{Z}$，每个子区域都有一个边缘计算服务器。边缘计算服务器表示为 $\mathcal{A}=\{a_1,\cdots,a_i,\cdots,a_M\}$，每个边缘计算服务器的能力相同，可服务相同计算量的计算任务。边缘计算服务器 a_i 的坐标表示为 (x_i,y_i)，边缘计算服务器 a_i 服务的覆盖范围表示为 $\mathcal{V}_i$，即边缘计算服务器 a_i 可为区域 $\mathcal{V}_i$ 中的物联网设备提供计算服务。当区域 $\mathcal{V}_i$ 中的物联网设备有计算任务需要卸载到边缘计算服务器进行处理时，物联网设备通过上行链路将计算任务卸载到边缘计算服务器 a_i 中进行处理。区域 $\mathcal{Z}$ 中物联网设备的数据总量为 N，且每个物联网设备需要卸载的数据量相同。卸载计算任务的物联网设备通过正交频分复用的方式接入边缘计算服务器。因此，当不同物联网设备将计算任务卸载到边缘计算服务器时，物联网设备之间不存在干扰。在物联网设备将计算任务卸载到边缘计算服务器的过程中，物联网设备的平均能耗、时延以及每个边缘计算服务器覆盖的物联网设备数量均与区域的分割方式密切相关。需要在满足边缘计算服务器覆盖的物联网设备卸载的计算量均衡前提下，合理规划边缘计算服务器的覆盖区域，以期尽可能降低物联网设备的平均能耗以及时延。

2.2 能耗最小化问题建模

考虑区域 $\mathcal{V}_i$ 中的物联网设备，坐标为 (x,y)，此设备将计算任务卸载到边缘计算服务器 a_i 中进行计算。该物联网设备到边缘计算服务器 a_i 的信道路径损耗可表示为

$$L_i(x,y)=\eta\left(\sqrt{(x-x_i)^2+(y-y_i)^2}\right)^{-\alpha} \tag{1}$$

其中，α 表示此信道的路径损耗指数，η 表示衰落因子。信道的可到达速率可表示为

$$R_i(x,y)=B\mathrm{lb}\left(1+\frac{P_i(x,y)L_i(x,y)}{N_0}\right) \tag{2}$$

其中，B 代表传输带宽，$P_i(x,y)$ 代表位于 (x,y) 的物联网设备卸载数据到边缘计算服务器 a_i 时的发射功率，N_0 表示噪声功率。为了满足可达速率为 r 的要求，物联网设备的最小发射功率可表示为

$$P_{\min,i}(x,y)=\frac{\left(2^{r/B}-1\right)N_0}{L_i(x,y)} \tag{3}$$

在区域 $\mathcal{Z}$ 中，物联网设备的分布服从二维分布 $f(x,y)$，所以在区域 $\mathcal{V}_i$ 中物联网设备的平均发射功率为

$$P_i=\iint_{\mathcal{V}_i}P_{\min,i}(x,y)f(x,y)\mathrm{d}x\mathrm{d}y \tag{4}$$

在确认每个边缘计算服务器的覆盖范围后，边缘计算服务器 a_i 覆盖的物联网设备数量可表示为

$$N_i=N\iint_{\mathcal{V}_i}f(x,y)\mathrm{d}x\mathrm{d}y \tag{5}$$

整个区域 $\mathcal{Z}$ 中物联网设备的平均发射功率为

$$P=\sum_{i=1}^{M}P_i=\sum_{i=1}^{M}\iint_{\mathcal{V}_i}P_{\min,i}(x,y)f(x,y)\mathrm{d}x\mathrm{d}y \tag{6}$$

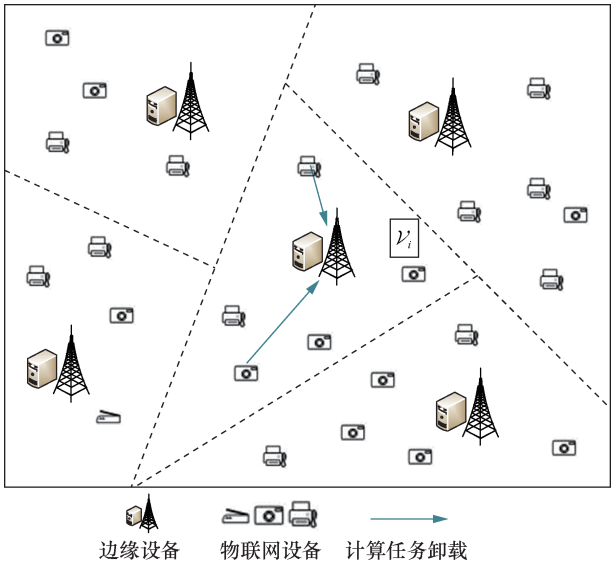

图 1 系统模型

本文的研究目标是在每个边缘计算服务器计算能力相同的情况下，确定边缘计算服务器覆盖范围的边界，在满足目标可达速率的情况下，使物联网设备平均发射功率尽可能小。功率最小化问题可表示为

$$\min_{\mathcal{V}_i}\sum_{i=1}^{M}\iint_{\mathcal{V}_i}\frac{\left(2^{r/B}-1\right)N_0}{L_i\left(x,y\right)}\cdot f(x,y)\mathrm{d}x\mathrm{d}y \tag{7a}$$

$$\text{s.t.}\quad \iint_{\mathcal{V}_i} f(x,y)\mathrm{d}x\mathrm{d}y=\omega_i,\quad \forall a_i\in\mathcal{A} \tag{7b}$$

$$\mathcal{V}_l\cap\mathcal{V}_m=\varnothing,\quad \forall l\neq m,\quad a_l,a_m\in\mathcal{A} \tag{7c}$$

$$\bigcup_{a_i\in\mathcal{A}}\mathcal{V}_i=\mathcal{Z} \tag{7d}$$

由于每个边缘计算服务器能力相同，因此 $\omega_i=\dfrac{1}{M},\forall a_i\in\mathcal{A}$。其中，式(7b)保证每个区域的负载约束，式(7c)和式(7d)保证每个区域是不相交的，并且所有区域的并集覆盖整个区域 $\mathcal{Z}$。

2.3　时延最小化问题建模

物联网设备将计算任务卸载到边缘计算服务器中进行计算，边缘计算服务器在完成计算任务后将结果传回物联网设备。本文采用正交频分复用的接入技术实现卸载过程，整个计算任务卸载过程的时延可表示为

$$t=t_{\mathrm{tx}}+t_{\mathrm{com}}+t_{\mathrm{rx}} \tag{8}$$

其中，t_{tx} 表示物联网设备卸载任务到边缘计算服务器过程中的传输时延，t_{com} 表示边缘计算服务器在计算过程中产生的时延，t_{rx} 表示边缘计算服务器将计算结果传输回物联网设备过程中产生的时延。由于 t_{com}、t_{rx} 与 t_{tx} 相比很小，可忽略不计，在本文中只考虑物联网设备卸载任务到边缘计算服务器过程中的传输时延 t_{tx} [5]。

本文每个物联网设备需卸载的任务量相同，均表示为 D，结合式(2)，传输时延可表示为

$$t_{\mathrm{tx}}=\frac{D}{B\mathrm{lb}\left(1+\dfrac{PL_i\left(x,y\right)}{N_0}\right)} \tag{9}$$

整个区域 $\mathcal{Z}$ 中的平均时延可表示为

$$T=\sum_{i=1}^{M}\iint_{\mathcal{V}_i}\frac{D}{B\mathrm{lb}\left(1+\dfrac{PL_i\left(x,y\right)}{N_0}\right)}\cdot f(x,y)\mathrm{d}x\mathrm{d}y \tag{10}$$

本文的另一个研究目标是在每个边缘计算服务器能力相同的情况下，确定边缘计算服务器覆盖范围的边界，以达到尽可能小的时延。因此，时延最小化问题可表示为

$$\min_{\mathcal{V}_i}\sum_{i=1}^{M}\iint_{\mathcal{V}_i}\frac{D}{B\mathrm{lb}\left(1+\dfrac{PL_i\left(x,y\right)}{N_0}\right)}\cdot f(x,y)\mathrm{d}x\mathrm{d}y \tag{11a}$$

$$\text{s.t.}\quad \iint_{\mathcal{V}_i} f(x,y)\mathrm{d}x\mathrm{d}y=\omega_i,\quad \forall a_i\in\mathcal{A} \tag{11b}$$

$$\mathcal{V}_l\cap\mathcal{V}_m=\varnothing,\quad \forall l\neq m,\quad a_l,a_m\in\mathcal{A} \tag{11c}$$

$$\bigcup_{a_i\in\mathcal{A}}\mathcal{V}_i=\mathcal{Z} \tag{11d}$$

在式(7)和式(11)的优化问题中，优化变量 $\mathcal{V}_i$ 是一系列连续的二维分区，并且相互影响，这使得优化问题的求解变得复杂。为了完美地描述用户的空间分布，将 $f(x,y)$ 视为 x 和 y 的泛型函数，这导致了给定的双重积分复杂性。另外，式(7b)和式(11b)的存在进一步加大了求解的难度。因此，直接求解式(7)和式(11)的优化问题是十分复杂的，本文利用最优运输理论[19]对其进行建模。

3　最优计算资源分配

3.1　最优运输理论

最优运输理论是本文解决能耗最小化以及时延最小化问题的主要理论基础。最优运输理论最初由 Monge 在处理运输问题时提出[19]，该理论主要用来研究两种概率分布之间的最优运输映射。Monge 问题的主要目的是寻找最优的运输方案，以最小的运输成本将一定量的沙子从一个区域运输到另外一个区域。

最优运输理论可理解为寻找两组集合之间的最优匹配，使集合之间的匹配总费用最小。两组集合可以是连续的也可以是离散的，并且可服从任意分布。最优运输理论中两个概率分布之间的匹配情况示意图如图 2 所示，$U\subset\mathbb{R}^n$ 和 $V\subset\mathbb{R}^n$ 均为已知可度量空间，$\mathbb{R}^n$ 代表 n 维空间，空间中的概率分布分别为 μ 和 ν，T 为从 U 到 V 的可能映射。接下来用数学语言表述 Monge 问题。已知在可度量空间 $U\subset\mathbb{R}^n$ 上的概率分布 μ 和可度量空间 $V\subset\mathbb{R}^n$ 中的概率分布 ν，单位费用函数为 $c\left(x,y\right)$，记为

$$\Gamma_1=\left\{T:U\to V;\mu\left(T^{-1}(E)\right)=\nu(E),\forall E\subseteq V\right\} \tag{12}$$

$$I(T)=\int c(x,T(x))\mathrm{d}\mu(x) \tag{13}$$

找到映射 T_0 使得

$$I\left(T_0\right)=\min_{T\in\Gamma_1} I(T) \tag{14}$$

其中，T_0 为最优映射，$I\left(T_0\right)$ 为最优运输费用。

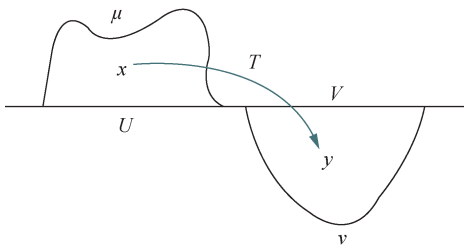

图 2 最优运输理论中两个概率分布之间的匹配情况示意图

Monge 问题是高度非线性的，而且要求源分布的点必须也只能映射到目的地的一个位置。Kantorovich 同样构造了最优运输问题，推广并松弛了 Monge 问题，使源分布的点可同时映射到目的地分布的多个点。松弛后的 Monge 问题被称为 Monge-Kantorovich 问题。Monge-Kantorovich 问题表述如下，已知可分度量空间 U,V 上的测度 μ,ν 和单位费用函数 $c(x,y)$，找到 $U\times V$ 上的概率测度 $\gamma_0\in\Gamma_2$，其中 Γ_2 表示上边际分布为 μ,ν 的概率测度全体，使满足

$$\min_{\gamma\in\Gamma_2}\int_{U\times V} c(x,y)\mathrm{d}\gamma(x,y) \tag{15}$$

称 γ_0 为最优运输计划或最优测度。将 Monge 问题松弛为 Monge-Kantorovich 问题后，可采用半连续的函数作为费用函数 $c(x,y)$。Monge-Kantorovich 问题是一个受线性约束的线性优化问题，因此可以利用对偶公式[20]求解最优解。Kantorovich 对偶原理可用数学语言表述。同 Monge-Kantorovich 问题类似，已知可分度量空间 U、V 上的测度 μ、ν，$c\left(x,y\right)$ 是一个下半连续的费用函数，根据对偶定理，式(15)等价于

$$\max_{\varphi,\psi}\left\{\int_U \psi\left(x\right)\mu\left(x\right)\mathrm{d}x+\int_V \varphi\left(y\right)\nu\left(y\right)\mathrm{d}y;\right.$$
$$\left.\psi(x)+\varphi(y)\leqslant c(x,y);x\in U,y\in V\right\} \tag{16}$$

其中，$\varphi\left(x\right)$ 和 $\psi\left(y\right)$ 为 Kantorovich 势函数，用于确定最优运输计划。Kantorovich 对偶定理为解决 Monge-Kantorovich 问题提供了一个可行的解决方案。但是，并不是在所有情况下 Monge 问题与 Monge-Kantorovich 问题的解都相同。在源分布和单位费用函数连续的情况下 Monge 问题与 Monge-Kantorovich 问题的解相同[21]，此时可将 Monge 问题转化为 Monge-Kantorovich 问题，并利用 Kantorovich 对偶定理进行求解。最优映射可表示为

$$T\left(x\right)=\left\{y\middle|\psi^*\left(x\right)+\varphi^*\left(y\right)=c\left(x,y\right)\right\} \tag{17}$$

其中，$\varphi^*\left(x\right)$ 和 $\psi^*\left(y\right)$ 为 Monge-Kantorovich 问题对偶公式对应的最优势函数。

本文的能耗最小化以及时延最小化问题可看作寻找数据在物联网设备以及边缘计算服务器之间运输的最优运输方案的问题。源分布为物联网设备的分布情况，而且源分布是一个连续的二元函数；目标分布是边缘计算服务器的位置；运输费用是数据运输过程中的能耗以及时延；最佳运输方案即为每个边缘计算服务器的覆盖范围。

3.2 最优计算资源分配

为了得到最优的边缘计算服务器覆盖范围，本文利用最优运输理论求解式(7)和式(11)。运输过程即数据从物联网设备传输到边缘计算服务器的过程。物联网设备的分布作为源分布且是连续的，边缘计算服务器即目标分布且是离散的点，传输过程中的平均能耗以及时延即费用函数，通过最优运输理论得到两个分布之间的最优映射。

在优化能耗的过程中，如式（7）所示，源分布为物联网设备的分布 $f(x,y)$，目的分布可表示为

$$\Lambda=\sum_{a_i\in\mathcal{A}}\omega_i\delta_{a_i} \tag{18}$$

其中，ω_i 与边缘计算服务器的工作能力相关，当所有边缘计算服务器可以服务的物联网设备数量相同时，$\omega_i=\dfrac{1}{M}$，δ_{a_i} 为狄拉克函数。传输过程中费用函数为

$$c_{\mathrm{E}}\left(x,y,a_i\right)=\frac{\left(2^{r/B}-1\right)N_0}{L_i\left(x,y\right)} \tag{19}$$

可以看出，源分布 $f(x,y)$ 与费用函数 $c_{\mathrm{E}}(x,y)$ 是连续的，此时 Monge 问题与 Monge-Kantorovich 问题的解相同，可将 Monge 问题松弛为 Monge-Kantorovich 问题并利用对偶公式求解

$$\min_{\gamma_i}\sum_{i=1}^{M}\iint_{\gamma_i}\frac{\left(2^{r/B}-1\right)N_0}{L_i\left(x,y\right)}\cdot f(x,y)\mathrm{d}x\mathrm{d}y=$$

$$\min_{\gamma_i}\sum_{i=1}^{M}\iint_{\gamma_i} c_{\mathrm{E}}\left(x,y,a_i\right)\cdot f(x,y)\mathrm{d}x\mathrm{d}y=$$

$$\max_{\psi,\varphi}\left\{\iint_{\mathcal{Z}}\psi(x,y)f(x,y)\mathrm{d}x\mathrm{d}y+\int_{\mathcal{A}}\varphi(a)\sum_{a_i\in\mathcal{A}}\omega_i\delta_{a-a_i}\mathrm{d}a;\ \psi(x,y)+\varphi(a)\leqslant c_{\mathrm{E}}(x,y,a_i)\right\}=$$
$$\max_{\psi,\varphi}\left\{\iint_{\mathcal{Z}}\psi(x,y)f(x,y)\mathrm{d}x\mathrm{d}y+\sum_{i=1}^{M}\varphi(a_i)\omega_i\ \psi(x,y)+\varphi(a_i)\leqslant c_{\mathrm{E}}(x,y,a_i),\forall a_i\in\mathcal{A}\right\}\tag{20}$$

为了使式(20)最大化，确定任意 φ、ψ 需要在可取范围内取最大值。又因为对于任意 $(x,y)\in\mathcal{Z}$，$a_i\in\mathcal{A}$，需要满足 $\psi(x,y)+\varphi(a_i)\leqslant c_{\mathrm{E}}(x,y,a_i)$。所以确定 φ 后，ψ 可表示为

$$\psi(x,y)=\varphi_{\mathrm{E}}(x,y)=\inf_i c_{\mathrm{E}}(x,y,a_i)-\varphi(a_i)\tag{21}$$

令 $\varphi_i=\varphi(a_i)$，结合式(20)和式(21)，可得出式(7)的优化问题等价于以下极大化问题

$$\max_{\varphi_i}\left\{F_{\mathrm{E}}(\varphi)=\iint_{\mathcal{Z}}\varphi_{\mathrm{E}}(x,y)f(x,y)\mathrm{d}x\mathrm{d}y+\sum_{i=1}^{M}\varphi_i\omega_i\right\}\tag{22a}$$

$$\varphi_{\mathrm{E}}(x,y)=\inf_i c_{\mathrm{E}}(x,y,a_i)-\varphi_i\tag{22b}$$

式(7)复杂的优化问题就转变为了式(22)有 M 个变量的易于处理的优化问题。通过解式(22)的优化问题将会得到一组 φ_i，通过这组 φ_i 可得到最优的边缘计算服务器覆盖范围。由式(17)可得，能耗最优的边缘计算服务器覆盖范围可表示为

$$\mathcal{V}_i^{\mathrm{E}}=\left\{(x,y)\middle|c_{\mathrm{E}}(x,y,a_i)-\varphi_i\leqslant c_{\mathrm{E}}(x,y,a_j)-\varphi_j,\forall j\neq i\right\}\tag{23}$$

可以通过计算 $F(\varphi)$ 的一阶导数，利用梯度下降法求解式(22)的优化问题。$F(\varphi)$ 的一阶导数可表示为

$$\frac{\partial F_{\mathrm{E}}}{\partial\varphi_i}=\varphi_i-\iint_{\mathcal{V}_i}f(x,y)\mathrm{d}x\mathrm{d}y\tag{24}$$

在优化时延的过程中，如式(11)所示，源分布为物联网设备的分布 $f(x,y)$，目的分布如式(18)所示，传输过程中费用函数为

$$c_{\mathrm{L}}(x,y,a_i)=\frac{D}{B\mathrm{lb}\left(1+\dfrac{PL_i(x,y)}{N_0}\right)}\tag{25}$$

可以看出，源分布 $f(x,y)$ 与费用函数 $c_{\mathrm{E}}(x,y)$ 是连续的，这时 Monge 问题与 Monge-Kantorovich 问题的解相同，可将 Monge 问题松弛为 Monge-Kantorovich 问题并利用对偶公式求解。式(11)的优化问题等价于以下极大化问题

$$\min_{\mathcal{V}_i}\sum_{i=1}^{M}\iint_{\mathcal{V}_i}\frac{D}{B\mathrm{lb}\left(1+\dfrac{PL_i(x,y)}{N_0}\right)}\cdot f(x,y)\mathrm{d}x\mathrm{d}y=$$
$$\max_{\varphi_i}\left\{F_{\mathrm{L}}(\varphi)=\iint_{\mathcal{Z}}\varphi_{\mathrm{L}}(x,y)f(x,y)\mathrm{d}x\mathrm{d}y+\sum_{i=1}^{M}\varphi_i\omega_i\right\}\tag{26a}$$

$$\varphi_{\mathrm{L}}(x,y)=\inf_i c_{\mathrm{L}}(x,y,a_i)-\varphi_i\tag{26b}$$

时延最优的边缘计算服务器覆盖范围可表示为

$$\mathcal{V}_i^{\mathrm{L}}=\left\{(x,y)\middle|c_{\mathrm{L}}(x,y,a_i)-\varphi_i\leqslant c_{\mathrm{L}}(x,y,a_j)-\varphi_j,\forall j\neq i\right\}\tag{27}$$

可以通过计算 $F(\varphi)$ 的一阶导数，利用梯度下降法求解式(26)的优化问题。

3.3　基于最优运输理论的资源分配机制

在物联网边缘计算场景中，可根据物联网设备的空间分布以及边缘计算服务器的位置信息，利用基于最优运输理论的资源分配机制，解决在边缘计算服务器计算能力有限且相同的约束下，计算资源的分配问题，实现能耗和时延的优化。能耗优化分配算法如算法 1 所示。在算法 1 中，$\delta>0$ 为停止算法的阈值。首先初始化数组 φ_t，利用式（24）计算 $\nabla F_E(\varphi_t)$，然后通过梯度下降法确定步长 ω_k，之后根据步长更新数组。满足条件 $\|\nabla F_E(\varphi_t)\|_2>\delta$ 时停止算法，并得到最优数组 φ_t，最后得到最优的边缘计算服务器覆盖范围 $\mathcal{V}_i^{\mathrm{E}}$，具体算法如下。

算法 1　能耗优化分配算法

输入　物联网设备分布 $f(x,y)$，阈值 δ，边缘计算服务器坐标 a_i，$\forall a_i\in\mathcal{A}$。

输出　边缘计算服务器覆盖范围 $\mathcal{V}_i^{\mathrm{E}}$

初始化数组 φ_t，$(t=1)$

while $\|\nabla F_{\mathrm{E}}(\varphi_t)\|_2>\delta$ do

　令 $k=1,\omega_1=1$

　更新 $\varphi_{t+1}=\varphi_t+\omega_1\nabla F_{\mathrm{E}}(\varphi_t)$

　if $\nabla F_{\mathrm{E}}(\varphi_t)<\nabla F_{\mathrm{E}}(\varphi_{t+1})$ then

　　while $\nabla F_{\mathrm{E}}(\varphi_t)<\nabla F_{\mathrm{E}}(\varphi_{t+1})$ do

　　$k\to k+1$

　　$\omega_k=2^{k-1}\omega_1$

　　更新 $\varphi_{t+1}=\varphi_t+\omega_k\nabla F_{\mathrm{E}}(\varphi_t)$

　end while

else

while $\nabla F_{\mathrm{E}}(\varphi_t) > \nabla F_{\mathrm{E}}(\varphi_{t+1})$ do

$k \to k+1$

$\omega_k = 2^{-k+1}\omega_1$

更新 $\varphi_{t+1} = \varphi_t + \omega_k \nabla F_{\mathrm{E}}(\varphi_t)$

end while

end if

$t \to t+1$

end while

$\varphi_i = \varphi_t(i)$

$\mathcal{V}_i^{\mathrm{E}} = \{(x,y) | c_{\mathrm{E}}(x,y,a_i) - \varphi_i \leqslant c_{\mathrm{E}}(x,y,a_j) - \varphi_j, \forall j \neq i\}$

时延优化分配算法如算法 2 所示。在算法 2 中，首先初始化数组 φ_t，然后通过梯度下降法确定步长 ω_k，之后根据步长更新数组。满足条件 $\|\nabla F_L(\varphi_t)\|_2 > \delta$ 时停止算法，并得到最优数组 φ_t，最后得到最优的边缘计算服务器覆盖范围 $\mathcal{V}_i^L$，具体算法如下。

算法 2 时延优化分配算法

输入 物联网设备分布 $f(x,y)$，阈值 δ，边缘计算服务器坐标 a_i，$\forall a_i \in \mathcal{A}$。

输出 边缘计算服务器覆盖范围 $\mathcal{V}_i^{\mathrm{L}}$

初始化数组 φ_t，$(t=1)$

while $\|\nabla F_{\mathrm{L}}(\varphi_t)\|_2 > \delta$ do

令 $k=1, \omega_1 = 1$

更新 $\varphi_{t+1} = \varphi_t + \omega_1 \nabla F_{\mathrm{L}}(\varphi_t)$

if $\nabla F_{\mathrm{L}}(\varphi_t) < \nabla F_{\mathrm{L}}(\varphi_{t+1})$ then

while $\nabla F_{\mathrm{L}}(\varphi_t) < \nabla F_{\mathrm{L}}(\varphi_{t+1})$ do

$k \to k+1$

$\omega_k = 2^{k-1}\omega_1$

更新 $\varphi_{t+1} = \varphi_t + \omega_k \nabla F_{\mathrm{L}}(\varphi_t)$

end while

else

while $\nabla F_{\mathrm{L}}(\varphi_t) > \nabla F_{\mathrm{L}}(\varphi_{t+1})$ do

$k \to k+1$

$\omega_k = 2^{-k+1}\omega_1$

更新 $\varphi_{t+1} = \varphi_t + \omega_k \nabla F_L(\varphi_t)$

end while

end if

$t \to t+1$

end while

$\varphi_i = \varphi_t(i)$

$\mathcal{V}_i^{\mathrm{L}} = \{(x,y) | c_{\mathrm{L}}(x,y,a_i) - \varphi_i \leqslant c_{\mathrm{L}}(x,y,a_j) - \varphi_j, \forall j \neq i\}$

综上所示，为解决在边缘计算服务器计算能力有限且相同的约束下，计算资源的分配问题，本文提出了一种基于最优运输理论的计算资源分配机制，基于 Kantorovich 对偶定理将物联网边缘计算系统中能耗和时延优化问题转化为可解的 Monge-Kantorovich 问题，进而利用梯度下降法给出了最优解，实现了能耗和时延的优化。

4 仿真分析

本节根据能耗优化、时延优化的两种区域划分机制进行仿真分析，并将本文提出的区域划分方法与泰森多边形划分机制以及随机分配法进行对比[17, 26-27]。

利用 MATLAB 工具进行仿真，在仿真过程中，本文考虑一个 1 000 m×1 000 m 的矩形区域，有 4 个边缘计算服务器分布其中。利用最优运输理论得到优化的分割机制时，物联网设备的分布可以采用任何合理的二元连续分布。在本文的仿真过程中，为模拟物联网设备分布不均匀的情况，物联网设备的分布采用二维的截断式高斯分布[21]，该分布适用于模拟存在热点的区域[28-29]，物联网设备分布情况如图 3 所示。

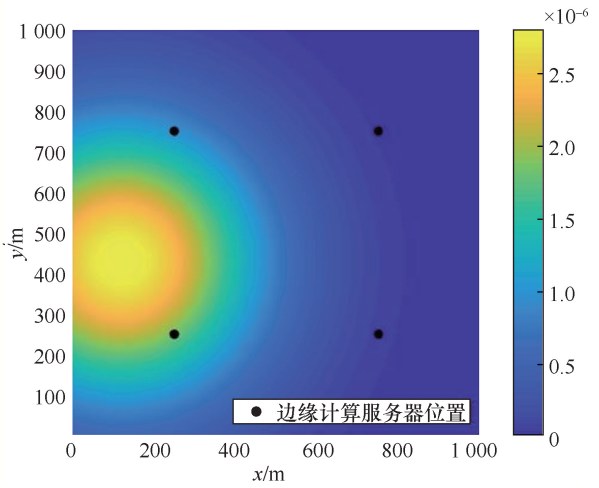

图 3 物联网设备分布情况

二维截断式高斯分布的表达式为[22]

$$f(x,y) = \frac{1}{G}\exp\left[-\left(\frac{x-\mu_x}{\sqrt{2}\sigma_x}\right)^2\right]\exp\left[-\left(\frac{y-\mu_y}{\sqrt{2}\sigma_y}\right)^2\right] \tag{28a}$$

$$G = 2\pi\sigma_x\sigma_y \operatorname{erf}\left(\frac{L_x-\mu_x}{\sqrt{2}\sigma_x}\right)\operatorname{erf}\left(\frac{L_y-\mu_y}{\sqrt{2}\sigma_y}\right) \tag{28b}$$

其中，L_x、L_y 分别表示区域的长宽，μ_x、μ_y 和 σ_x、σ_y 分别代表 x 和 y 的平均值和标准差。另外，$\mathrm{erf}(z)=\frac{2}{\sqrt{\pi}}\int_0^z \mathrm{e}^{-t^2}\mathrm{d}t$。在上述分布中，热点坐标为 (μ_x,μ_y)，由 σ_x 和 σ_y 确定热点附近物联网设备密度，并且热点附近物联网设备密度与 σ_x 和 σ_y 的值成反比。定义 $\rho_x=\frac{1}{\sigma_x}$ 和 $\rho_y=\frac{1}{\sigma_y}$ 分别为热点坐标处 x 方向和 y 方向的用户密度，且 $\rho_x=\rho_y=\rho$[25]。仿真参数如表 1 所示。

表 1　　仿真参数

参数	说明	数值
L_x、L_y	区域边长	1 000 m
M	边缘计算服务器数	4
(x_i,y_i)	边缘计算服务器坐标	（250,250）（750,250）（750,750）（250,750）
N	物联网设备数量	600
D	物联网设备需卸载数据量	100 MB
μ_x、μ_y	二维截断式高斯分布平均值	120 m、430 m

边缘计算服务器服务均衡性对比如图 4 所示，在边缘计算服务器服务能力相同的情况下，利用最优运输理论得到的能耗优化分配机制、时延优化分配机制以及随机分配机制，可以很好地保证边缘计算服务器的服务均衡性，即每个边缘计算服务器服务的物联网设备数量相同。采用泰森多边形划分机制时，由于不考虑物联网设备的空间分布情况，每个物联网设备选择距离最近的边缘计算服务器，每个边缘计算服务器服务的物联网设备数量相差巨大。在物联网设备分布密度较大的区域，边缘计算服务器由于能力有限而不能完成所有任务，在物联网设备分布密度较小的区域，将会出现计算资源的浪费。与经典的泰森多边形划分机制相比，利用最优运输理论得到的能耗优化机制与时延优化机制均拥有更好的均衡性，每个边缘计算服务器服务的物联网设备数量基本相同，有效避免了出现过载或者资源浪费的情况。

能耗优化分割机制下物联网设备的平均功率与带宽关系如图 5 所示。针对基于最优运输理论的能耗优化分割机制，在边缘计算服务器数量一定的情况下，随着带宽的增大，适当减少发射速率同样

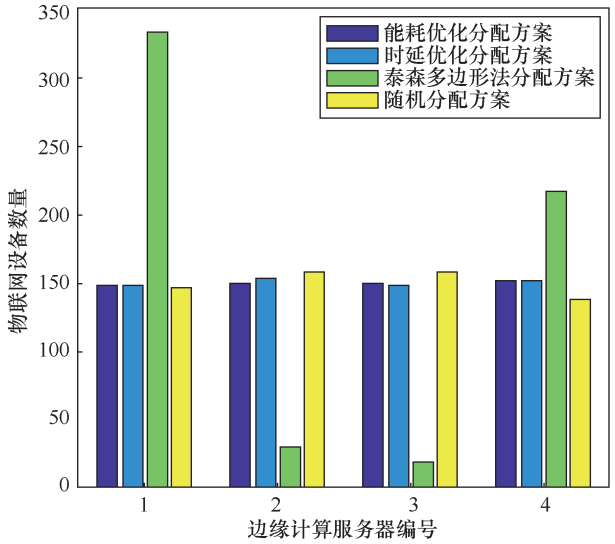

图 4　边缘计算服务器服务均衡性对比

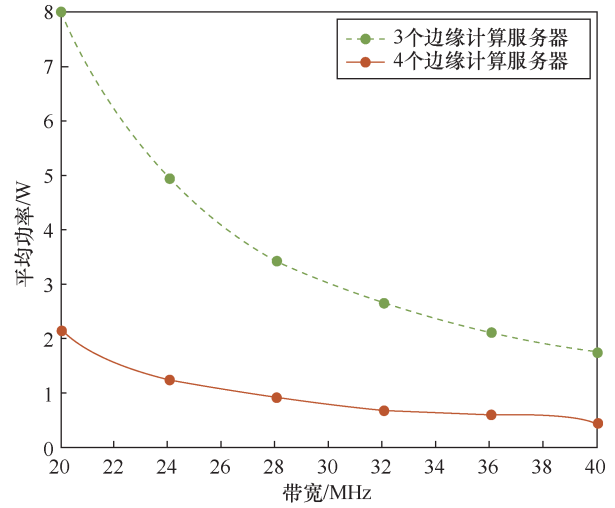

图 5　能耗优化分割机制下物联网设备的平均功率与带宽关系

可以达到要求的传输速率，因此物联网设备的平均发射功率随带宽的增大而减小。在带宽一定的情况下，随着边缘计算服务器的增加，物联网设备的平均发射功率减小。

在不同分配机制下，物联网设备的平均功率与带宽关系如图 6 所示，在随机分配机制、加权泰森多边形划分机制以及能耗优化分配机制下，物联网设备的平均发射功率均随带宽的增大而降低。在带宽一定的情况下，利用最优运输理论得到的能耗优化分割机制比采用随机分配机制达到的平均发射功率更小，并且平均发射功率最多可降低 24%。在带宽一定的情况下，利用最优运输理论得到的能耗优化分割机制比采用加权泰森多边形划分机制达到的平均发射功率更小，并且平均发射功率最多可降低 21%。

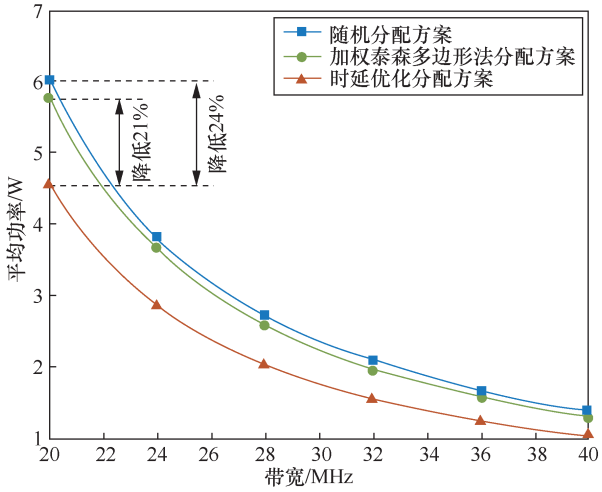

图 6　物联网设备的平均功率与带宽关系

在不同机制下，物联网设备平均时延与带宽关系如图 7 所示，在同一区域，采用随机分配机制、加权泰森多边形划分机制以及时延优化分配机制时，物联网设备的平均时延均随带宽的增大而减小。在带宽一定的情况下，利用最优运输理论得到的时延优化分割机制比采用随机分配机制达到的平均时延更小，并且平均时延最多可降低 51%。在带宽一定的情况下，利用最优运输理论得到的时延优化分割机制比采用加权泰森多边形划分机制达到的平均时延更小，并且平均时延最多可降低 45%。

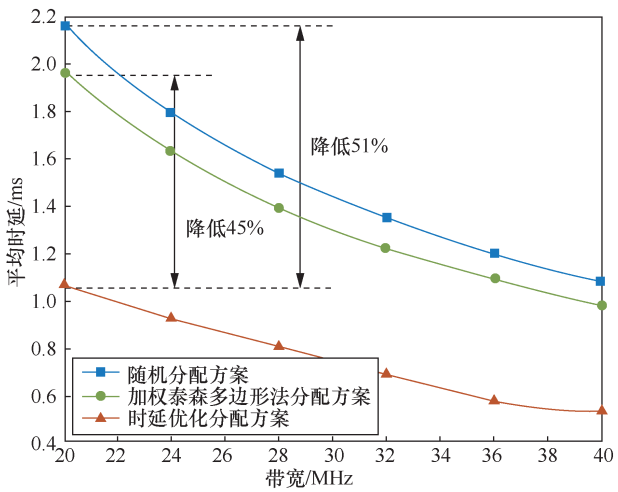

图 7　物联网设备平均时延与带宽关系

在不同机制下，物联网设备平均时延与物联网设备热点处密度关系如图 8 所示。在同一区域，热点处物联网设备密度越大，越多的用户集中在热点附近并且将计算任务卸载到距离最近的同一边缘计算服务器上，采用加权泰森多边形划分机制以及时延优化分配机制时，物联网设备的平均时延均随物联网设备热点处密度的增大而增大。在物联网设备热点处密度一定的情况下，利用最优运输理论得到的时延优化分割机制比采用随机分配机制达到的平均时延更小，并且 $\rho=0.02$ 时平均时延可降低 24%。在物联网设备热点处密度一定的情况下，利用最优运输理论得到的时延优化分割机制比采用加权泰森多边形划分机制达到的平均时延更小，并且 $\rho=0.02$ 时平均时延可降低 22%。

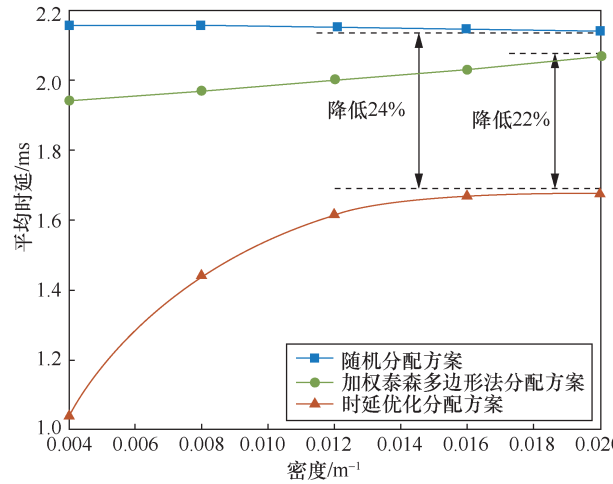

图 8　物联网设备平均时延与物联网设备热点处密度关系

在不同机制下，物联网设备平均功率与物联网设备热点处密度关系如图 9 所示。在同一区域，热点处物联网设备密度越大，越多的用户集中在热点附近，由于存在边缘计算服务器计算能力相同的约束，一部分物联设备需要将计算任务卸载到较远的边缘计算服务器中，采用时延优化分配机制时，物联网设备的平均功率均随物联网设备热点处密度的增大而增大。在物联网设备热点处密度一定的情况下，利用最优运输理论得到的时延优化分割机制比采用随机分配机制达到的平均功率更小，并且 $\rho=0.024$ 时平均功率可降低 19%。在物联网设备热点处密度一定的情况下，利用最优运输理论得到的时延优化分割机制比采用加权泰森多边形划分机制达到的平均功率更小，并且 $\rho=0.024$ 时平均功率可降低 14%。

5　结束语

本文利用最优运输理论进行计算资源分配，考虑边缘计算服务器计算能力的约束，针对物联网设备的能耗以及时延性能优化，提出了一种基于物联

网设备分布和边缘计算服务器位置的地理区域优化划分机制。仿真结果表明，与传统泰森多边形划分机制相比，本文提出的优化机制具有更好的均衡性，并且物联网设备的平均能耗最多降低 21%，平均时延最多降低 45%。在未来的工作中，将在基于最优运输理论的物联网边缘计算资源分配机制的基础上进行能耗和时延之间权衡的研究。

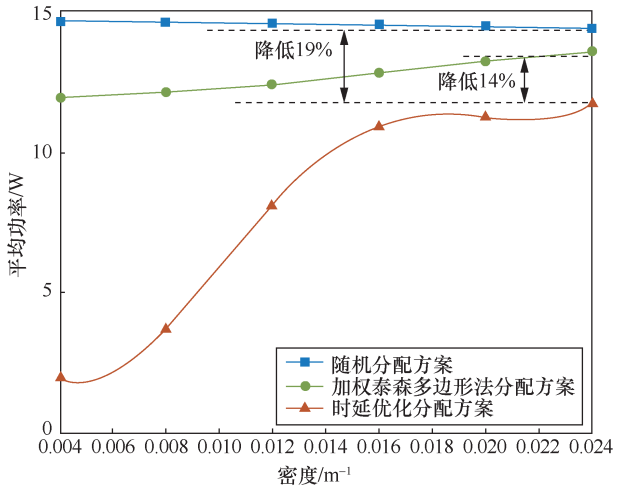

图 9　物联网设备平均功率与物联网设备热点处密度关系

参考文献：

[1] SHARMA S K, WANG X B. Live data analytics with collaborative edge and cloud processing in wireless IoT networks[J]. IEEE Access, 2017(5): 4621-4635.

[2] JIANG Y N, GE X H, ZHONG Y, et al. A new small-world IoT routing mechanism based on cayley graphs[J]. IEEE Internet of Things Journal, 2019, 6(6): 10384 - 10395.

[3] LIU H, ELDARRAT F, ALQAHTANI H, et al. Mobile edge cloud system: architectures, challenges, and approaches[J]. IEEE Systems Journal, 2018, 12(3): 2495-2508.

[4] ZHANG W W, WEN Y G, GUAN K, et al. Energy-optimal mobile cloud computing under stochastic wireless channel[J]. IEEE Transactions on Wireless Communications, 2013, 12(9): 4569-4581.

[5] YOU C S, HUANG K B, CHAE H, et al. Energy-efficient resource allocation for mobile-edge computation offloading[J]. IEEE Transactions on Wireless Communications, 2017, 16(3): 1397-1411.

[6] CHEN X, JIAO L, LI W Z, et al. Efficient multi-user computation offloading for mobile-edge cloud computing[J]. IEEE/ACM Transactions on Networking, 2016, 24(5): 2795-2808.

[7] SARDELLITTI S, SCUTARI G, BARBAROSSA S. Joint optimization of radio and computational resources for multicell mobile-edge computing[J]. IEEE Transactions on Signal and Information Processing Over Networks, 2015, 1(2): 89-103.

[8] ZHAO T C, ZHOU S, GUO X Y, et al. A cooperative scheduling scheme of local cloud and Internet cloud for delay-aware mobile cloud computing[C]//Proceedings of 2015 IEEE Globecom Workshops (GC Wkshps). Piscataway: IEEE Press, 2015: 1-6.

[9] LIU J, MAO Y Y, ZHANG J, et al. Delay-optimal computation task scheduling for mobile-edge computing systems[C]//Proceedings of 2016 IEEE International Symposium on Information Theory (ISIT). Piscataway: IEEE Press, 2016: 1451-1455.

[10] XU J L, PALANISAMY B, LUDWIG H, et al. Zenith: utility-aware resource allocation for edge computing[C]//Proceedings of 2017 IEEE International Conference on Edge Computing (EDGE). Piscataway: IEEE Press, 2017: 47-54.

[11] KWAK J, KIM Y, LEE J, et al. DREAM: dynamic resource and task allocation for energy minimization in mobile cloud systems[J]. IEEE Journal on Selected Areas in Communications, 2015, 33(12): 2510-2523.

[12] JIANG Z F, MAO S W. Energy delay trade-off in cloud offloading for mutli-core mobile devices[C]//Proceedings of 2015 IEEE Global Communications Conference (GLOBECOM). Piscataway: IEEE Press, 2015: 1-6.

[13] MUÑOZ O, PASCUAL-ISERTE A, VIDAL J. Optimization of radio and computational resources for energy efficiency in latency-constrained application offloading[J]. IEEE Transactions on Vehicular Technology, 2015, 64(10): 4738-4755.

[14] ZHAO J H, LI Q P, GONG Y, et al. Computation offloading and resource allocation for cloud assisted mobile edge computing in vehicular networks[J]. IEEE Transactions on Vehicular Technology, 2019, 68(8): 7944-7956.

[15] DENG R L, LU R X, LAI C Z, et al. Optimal workload allocation in fog-cloud computing toward balanced delay and power consumption[J]. IEEE Internet of Things Journal, 2016, 3(6): 1171-1181.

[16] CHEN X F, ZHANG H G, WU C, et al. Optimized computation offloading performance in virtual edge computing systems via deep reinforcement learning[J]. IEEE Internet of Things Journal, 2019, 6(3): 4005-4018.

[17] CHENG N, LYU F, QUAN W, et al. Space/aerial-assisted computing offloading for IoT applications: a learning-based approach[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(5): 1117-1129.

[18] WANG J D, ZHAO L, LIU J J, et al. Smart resource allocation for mobile edge computing: a deep reinforcement learning approach[J]. IEEE Transactions on Emerging Topics in Computing, 2661(99): 1.

[19] VILLANI C. Topics in optimal transportation[M]. Providence, Rhode Island: American Mathematical Society, 2003.

[20] SANTAMBROGIO F. Optimal transport for applied mathematicians[M]. Cham: Springer International Publishing, 2015.

[21] AMBROSIO L, GIGLI N. A user's guide to optimal transport[M]. Berlin Heidelberg: Springer, 2013.

[22] GHAZZAI H. Environment aware cellular networks[D]. Thuwal: King Abdullah University of Science and Technology, 2015.

[23] GHAZZAI H, TEMBINE H, ALOUINI M S. Mobile user association for heterogeneous networks using optimal transport theory[C]//Proceedings of 2017 6th International Conference on Communications and Networking (ComNet). Piscataway: IEEE Press, 2017: 1-6.

[24] WANG Y, HU Z Q, WEN X M, et al. Three-dimensional aerial cell partitioning based on optimal transport theory[C]//Proceedings of 2020 IEEE International Conference on Communications Workshops (ICC Workshops). Piscataway: IEEE Press, 2020: 1-6.

[25] MOZAFFARI M, SAAD W, BENNIS M, et al. Optimal transport

theory for power-efficient deployment of unmanned aerial vehicles[C]//Proceedings of 2016 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2016: 1-6.

[26] MOZAFFARI M, SAAD W, BENNIS M, et al. Wireless communication using unmanned aerial vehicles (UAVs): optimal transport theory for hover time optimization[J]. IEEE Transactions on Wireless Communications, 2017, 16(12): 8052-8066.

[27] GE X H, YANG B, YE J L, et al. Spatial spectrum and energy efficiency of random cellular networks[J]. IEEE Transactions on Communications, 2015, 63(3): 1019-1030.

[28] GE X H, YE J L, YANG Y, et al. User mobility evaluation for 5G small cell networks based on individual mobility model[J]. IEEE Journal on Selected Areas in Communications, 2016, 34(3): 528-541.

[29] GE X H, TU S, MAO G Q, et al. 5G ultra-dense cellular networks[J]. IEEE Wireless Communications, 2016, 23(1): 72-79.

[作者简介]

张琪（1997− ），女，华中科技大学硕士生，主要研究方向为无线通信、环境智适应网络柔性传输理论和边缘计算。

蒋宇娜（1994− ），女，华中科技大学博士生，主要研究方向为无线通信、区块链和物联网。

葛晓虎（1972− ），男，博士，华中科技大学教授，主要研究方向为移动通信、无线网络中的流量建模、绿色通信等。

李永会（1975− ），男，博士，澳大利亚悉尼大学教授，主要研究方向为无线通信、物联网、无线 AI 等。